file caplus
COST IN U.S. DOLLARS

	SINCE FILE ENTRY	TOTAL SESSION
FULL ESTIMATED COST	0.21	0.21

FILE 'CAPLUS' ENTERED AT 14:50:08 ON 04 MAY 2004
USE IS SUBJECT TO THE TERMS OF YOUR STN CUSTOMER AGREEMENT.
PLEASE SEE "HELP USAGETERMS" FOR DETAILS.

FILE COVERS 1907 - 4 May 2004 VOL 140 ISS 19
FILE LAST UPDATED: 3 May 2004 (20040503/ED)

This file contains CAS Registry Numbers for easy and accurate substance identification.

```
=> s (protein or peptide or polypeptide) and alignment and score and gap and matrix
       1618733 PROTEIN
       1116094 PROTEINS
       1875900 PROTEIN
                 (PROTEIN OR PROTEINS)
        309153 PEPTIDE
        225983 PEPTIDES
        395652 PEPTIDE
                 (PEPTIDE OR PEPTIDES)
         94004 POLYPEPTIDE
         54516 POLYPEPTIDES
        128177 POLYPEPTIDE
                 (POLYPEPTIDE OR POLYPEPTIDES)
         51734 ALIGNMENT
          5043 ALIGNMENTS
         54839 ALIGNMENT
                 (ALIGNMENT OR ALIGNMENTS)
         20349 SCORE
         18614 SCORES
         34891 SCORE
                 (SCORE OR SCORES)
        163496 GAP
         26329 GAPS
        179257 GAP
                 (GAP OR GAPS)
        419184 MATRIX
         56979 MATRIXES
          7789 MATRICES
        448622 MATRIX
                 (MATRIX OR MATRIXES OR MATRICES)
L1          23 (PROTEIN OR PEPTIDE OR POLYPEPTIDE) AND ALIGNMENT AND SCORE AND
               GAP AND MATRIX

=> d bib,abs 1-23

L1   ANSWER 1 OF 23  CAPLUS  COPYRIGHT 2004 ACS on STN
```

AN 2004:296101 CAPLUS
TI Program on pair wise sequence **alignment**
AU Sharad, Pranav; Jaipuriar, Sumeet
CS IInd Year B. Tech, Bioinformatics, Vellore Institute of Technology, Vellore, 632014, India
SO Bioinformatics India (2003), 1(3), 39-42
CODEN: BIINGI; ISSN: 0972-7655
PB Bioinformatics Institute of India
DT Journal
LA English
AB This program performs pair wise **alignment** of **protein** sequences and sequences that code for **proteins** DNA. It performs local as well as global **alignment** using BLOSUM-50 and PAM-250 scoring **matrixes** as per user's choice. The user also views the **alignment** with **gaps** and the **score** of the **alignment**. It also gives the percentage **alignment** of the sequences, which helps in determining the structural functional and evolutionary relationship of the sequences. The program uses Needleman-Wunsch algorithm for global **alignment** and Smith-Waterman algorithm for global **alignment** and local **alignment** resp., to obtain the dynamic programming **matrix** . The program is coded in c++ and is compiled under windows.
RE.CNT 3 THERE ARE 3 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 2 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 2002:671022 CAPLUS
DN 137:334852
TI Optimization of a new **score** function for the generation of accurate **alignments**
AU Qian, Bin; Goldstein, Richard A.
CS Biophysics Research Division, University of Michigan, Ann Arbor, MI, 48109-1055, USA
SO Proteins: Structure, Function, and Genetics (2002), 48(4), 605-610
CODEN: PSFGEY; ISSN: 0887-3585
PB Wiley-Liss, Inc.
DT Journal
LA English
AB The accuracy of the **alignments** of **protein** sequences depends on the **score matrix** and **gap** penalties used in performing the **alignment**. Most **score** functions are designed to find homologs in the various databases rather than to generate accurate **alignments** between known homologs. We describe the optimization of a **score** function for the purpose of generating accurate **alignments**, as evaluated by using a coordinate root-mean-square deviation (RMSD)-based merit function. We show that the resulting **score matrix**, which we call STROMA, generates more accurate **alignments** than other commonly used **score matrixes**, and this difference is not due to differences in the **gap** penalties. In fact, in contrast to most of the other **matrixes**, the **alignment** accuracies with STROMA are relatively insensitive to the choice of **gap** penalty parameters.
RE.CNT 34 THERE ARE 34 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 3 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 2002:527696 CAPLUS
DN 138:13797
TI Sequence **alignment** and database searching
AU Schuler, Gregory D.
CS National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
SO Methods of Biochemical Analysis (2001), 43(Bioinformatics, (2nd Edition)),

187-214
CODEN: MBANAA; ISSN: 0076-6941
PB John Wiley & Sons, Inc.
DT Journal; General Review
LA English
AB A review decribes some of the fundamental concepts involved in sequence **alignment**, particularly pairwise **alignments**, and database searching. Topics covered include the evolutionary basis of sequence **alignment**; modular nature of **proteins**; optimal **alignment** methods; substitution **scores** and **gap** penalties; statistical significance of **alignments**; database similarity searching; FASTA and BLAST programs; database searching artifacts; position-specific scoring **matrixes**; and spliced **alignments**.
RE.CNT 47 THERE ARE 47 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 4 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 2002:226437 CAPLUS
DN 137:104338
TI The efficient computation of position-specific match **scores** with the fast fourier transform
AU Rajasekaran, S.; Jin, X.; Spouge, J. L.
CS Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, 32611, USA
SO Journal of Computational Biology (2002), 9(1), 23-33
CODEN: JCOBEM; ISSN: 1066-5277
PB Mary Ann Liebert, Inc.
DT Journal
LA English
AB Historically, in computational biol. the fast Fourier transform (FFT) has been used almost exclusively to count the number of exact letter matches between two biosequences. This paper presents an FFT algorithm that can compute the match **score** of a sequence against a position-specific scoring **matrix** (PSSM). Our algorithm finds the PSSM **score** simultaneously over all offsets of the PSSM with the sequence, although like all previous FFT algorithms, it still disallows **gaps**. Although our algorithm is presented in the context of global matching, it can be adapted to local matching without **gaps**. As a benchmark, our PSSM-modified FFT algorithm computed pairwise match **scores**. In timing expts., our most efficient FFT implementation for pairwise scoring appeared to be 10 to 26 times faster than a traditional FFT implementation, with only a factor of 2 in the acceleration attributable to a previously known compression scheme. Many important algorithms for detecting biosequence similarities, e.g., gapped BLAST or PSIBLAST, have a heuristic screening phase that disallows **gaps**. This paper demonstrates that FFT algorithms merit reconsideration in these screening applications.
RE.CNT 35 THERE ARE 35 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 5 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 2002:60461 CAPLUS
DN 137:197786
TI Estimation of P-values for global **alignments** of **protein** sequences
AU Webber, Caleb; Barton, Geoffrey J.
CS EMBL-European Bioinformatics Institute, Hinxton, CB10 1SD, UK
SO Bioinformatics (2001), 17(12), 1158-1167
CODEN: BOINFP; ISSN: 1367-4803
PB Oxford University Press
DT Journal
LA English
AB The global **alignment** of **protein** sequence pairs is

often used in the classification and anal. of full-length sequences. The calcn. of a Z-**score** for the comparison gives a length and composition corrected measure of the similarity between the sequences. However, the Z-**score** alone, does not indicate the likely biol. significance of the similarity. In this paper, all pairs of domains from 250 sequences belonging to different SCOP folds were aligned and Z-**scores** calculated The distribution of Z-**scores** was fitted with a peak distribution from which the probability of obtaining a given Z-**score** from the global **alignment** of two **protein** sequences of unrelated fold was calculated A similar anal. was applied to subsequence pairs found by the Smith-Waterman algorithm. These analyses allow the probability that two **protein** sequences share the same fold to be estimated by global sequence **alignment**. The relationship between Z-**score** and probability varied little over the **matrix/gap** penalty combinations examined However, an average shift of +4.7 was observed for Z-**scores** derived from global **alignment** of locally-aligned subsequences compared to global **alignment** of the full-length sequences. This shift was shown to be the result of pre-selection by local **alignment**, rather than any structural similarity in the subsequences. The search ability of both methods was benchmarked against the SCOP superfamily classification and showed that global **alignment** Z-**scores** generated from the entire sequence are as effective as SSEARCH at low error rates and more effective at higher error rates. However, global **alignment** Z-**scores** generated from the best locallyaligned subsequence were significantly less effective than SSEARCH. The method of estimating statistical significance described here was shown to give similar values to SSEARCH and BLAST, providing confidence in the significance estimation

RE.CNT 38 THERE ARE 38 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 6 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 2001:925408 CAPLUS
DN 136:364379
TI Computational complexity of multiple sequence **alignment** with SP-**score**
AU Just, Winfried
CS Department of Mathematics, College of Arts & Sciences, Ohio University, Athens, OH, 45701, USA
SO Journal of Computational Biology (2001), 8(6), 615-623
CODEN: JCOBEM; ISSN: 1066-5277
PB Mary Ann Liebert, Inc.
DT Journal
LA English
AB It is shown that the multiple **alignment** problem with SP-**score** is NP-hard for each scoring **matrix** in a broad class M that includes most scoring **matrixes** actually used in biol. applications. The problem remains NP-hard even if sequences can only be shifted relative to each other and no internal **gaps** are allowed. It is also shown that there is a scoring **matrix** M0 such that the multiple **alignment** problem for M0 is MAX-SNP-hard, regardless of whether or not internal **gaps** are allowed.

RE.CNT 18 THERE ARE 18 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 7 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 2001:716031 CAPLUS
DN 136:352232
TI AL2CO: Calculation of positional conservation in a **protein** sequence **alignment**
AU Pei, Jimin; Grishin, Nick V.
CS Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX, 75390-9050, USA
SO Bioinformatics (2001), 17(8), 700-712

CODEN: BOINFP; ISSN: 1367-4803
PB Oxford University Press
DT Journal
LA English
AB Amino acid sequence **alignments** are widely used in the anal. of **protein** structure, function and evolutionary relationships. **Proteins** within a superfamily usually share the same fold and possess related functions. These structural and functional constraints are reflected in the **alignment** conservation patterns. Positions of functional and/or structural importance tend to be more conserved. Conserved positions are usually clustered in distinct motifs surrounded by sequence segments of low conservation. Poorly conserved regions might also arise from the imperfections in multiple **alignment** algorithms and thus indicate possible **alignment** errors. Quantification of conservation by attributing a conservation index to each aligned position makes motif detection more convenient. Mapping these conservation indexes onto a **protein** spatial structure helps to visualize spatial conservation features of the mol. and to predict functionally and/or structurally important sites. Anal. of conservation indexes could be a useful tool in detection of potentially misaligned regions and will aid in improvement of multiple **alignments**. We developed a program to calculate a conservation index at each position in a multiple sequence **alignment** using several methods. Namely, amino acid frequencies at each position are estimated and the conservation index is calculated from these frequencies. We utilize both unweighted frequencies and frequencies weighted using two different strategies. Three conceptually different approaches (entropy-based, variance-based and **matrix score**-based) are implemented in the algorithm to define the conservation index. Calculating conservation indexes for 35522 positions in 284 **alignments** from SMART database we demonstrate that different methods result in highly correlated (correlation coefficient more than 0.85) conservation indexes. Conservation indexes show statistically significant correlation between sequentially adjacent positions i and i + j, where j < 13, and averaging of the indexes over the window of three positions is optimal for motif detection. Positions with **gaps** display substantially lower conservation properties. We compare conservation properties of the SMART **alignments** or FSSP structural **alignments** to those of the ClustalW **alignments**. The results suggest that conservation indexes should be a valuable tool of **alignment** quality assessment and might be used as an objective function for refinement of multiple **alignments**.
RE.CNT 54 THERE ARE 54 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 8 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 2001:285733 CAPLUS
DN 136:48842
TI Making sense of **score** statistics for sequence **alignments**
AU Pagni, Marco; Jongeneel, C. Victor
CS Swiss Institute of Bioinformatics, Switz.
SO Briefings in Bioinformatics (2001), 2(1), 51-67
CODEN: BBIMFX; ISSN: 1467-5463
PB Henry Stewart Publications
DT Journal; General Review
LA English
AB A review and discussion. The search for similarity between two biol. sequences lies at the core of many applications in bioinformatics. This paper aims to highlight a few of the principles that should be kept in mind when evaluating the statistical significance of **alignments** between sequences. The extreme value distribution is first introduced, which in most cases describes the distribution of **alignment** **scores** between a query and a database. The effects of the similarity **matrix** and **gap** penalty values on the

score distribution are then examined, and it is shown that the **alignment** statistics can undergo an abrupt phase transition. A few types of random sequence databases used in the estimation of statistical significance are presented, and the statistics employed by the BLAST, FASTA and PRSS programs are compared. Finally the different strategies used to assess the statistical significance of the matches produced by profiles and hidden Markov models are presented.

RE.CNT 25 THERE ARE 25 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 9 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 2000:456393 CAPLUS
DN 133:249265
TI Accurate Formula for P-Values of Gapped Local Sequence and Profile **Alignments**
AU Mott, Richard
CS Wellcome Trust Centre for Human Genetics, Oxford, OX3 7BN, UK
SO Journal of Molecular Biology (2000), 300(3), 649-659
CODEN: JMOBAK; ISSN: 0022-2836
PB Academic Press
DT Journal
LA English
AB A simple general approximation for the distribution of gapped local **alignment scores** is presented, suitable for assessing significance of comparisons between two **protein** sequences or a sequence and a profile. The approximation takes account of the scoring scheme (i.e. **gap** penalty and substitution **matrix** or profile), sequence composition and length. Use of this formula means it is unnecessary to fit an extreme-value distribution to simulations or to the results of databank searches. The method is based on the theor. ideas introduced by R. Mott and R. Tribe in 1999. Extensive simulation studies show that **score**-thresholds produced by the method are accurate to within ±5% 95 % of the time. We also investigate factors which effect the accuracy of **alignment** statistics, and show that any method based on asymptotic theory is limited because asymptotic behavior is not strictly achieved for many real **protein** sequences, due to extreme composition effects. Consequently, it may not be practicable to find a general formula that is significantly more accurate until the sub-asymptotic behavior of **alignments** is better understood. (c) 2000 Academic Press.

RE.CNT 37 THERE ARE 37 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 10 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1999:234978 CAPLUS
DN 131:16078
TI Approximate statistics of gapped **alignments**
AU Mott, Richard; Tribe, Roger
CS Wellcome Trust Cent. Human Genetics, Oxford, UK
SO Journal of Computational Biology (1999), 6(1), 91-112
CODEN: JCOBEM; ISSN: 1066-5277
PB Mary Ann Liebert, Inc.
DT Journal
LA English
AB A heuristic approximation to the **score** distribution of gapped **alignments** in the logarithmic domain is presented. The method applies to comparisons between random, unrelated **protein** sequences, using standard **score matrixes** and arbitrary **gap** penalties. It is shown that gapped **alignment** behavior is essentially governed by a single parameter, α, depending on the penalty scheme and sequence composition This treatment also predicts the position of the transition point between logarithmic and linear behavior. The approximation is tested by simulation and shown to be accurate over a range of commonly used substitution **matrixes** and

gap-penalties.
RE.CNT 38 THERE ARE 38 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 11 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1998:755874 CAPLUS
DN 130:206894
TI Multiple model approach: Exploring the limits of comparative modeling
AU Jaroszewski, Lukasz; Pawlowski, Krzysztof; Godzik, Adam
CS Department of Chemistry, University of Warsaw, Warsaw, Pol.
SO Journal of Molecular Modeling [Electronic Publication] (1998), 4(10), 294-309
CODEN: JMMOFK; ISSN: 0948-5023
URL: http://link.springer.de/link/service/journals/00894/papers//80040010/80040294.pdf
PB Springer-Verlag
DT Journal; (online computer file)
LA English
AB One of the biggest problems in modeling distantly related **proteins** is the quality of the target-template **alignment**. This problem often results in low quality models that do not utilize all the information available in the template structure. The divergence of **alignments** at a low sequence identity level, which is a hindrance in most modeling attempts, is used here as a basis for a new technique of Multiple Model Approach (MMA). Alternative **alignments** prepared here using different mutation **matrixes** and **gap** penalties, combined with automated model building, are used to create a set of models that explore a range of possible conformations for the target **protein**. Models are evaluated using different techniques to identify the best model. In the set of examples studied here, the correct target structure is known, which allows the evaluation of various **alignment** and evaluation strategies. For a randomly selected group of distantly homologous **protein** pairs representing all structural classes and various fold types, it is shown that a threading **score** based on simplified statistical potentials of mean force can identify the best models and, consequently, the most reliable **alignment**. In cases where the difference between target and template structures is significant, the threading **score** shows clearly that all models are wrong, therefore disqualifying the template.

L1 ANSWER 12 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1998:610031 CAPLUS
DN 130:1689
TI Clusterization of P450 superfamily using the objective pair **alignment** method and the UPGMA program
AU Archakov, Alexander I.; Lisitsa, Andrey V.; Zgoda, Victor G.; Ivanova, Marina S.; Koymans, Luc
CS Inst. of Biomed. Chem., Moscow, 119832, Russia
SO Journal of Molecular Modeling [Electronic Publication] (1998), 4(7), 234-238
CODEN: JMMOFK; ISSN: 0948-5023
URL: http://link.springer.de/link/service/journals/00894/papers/8004007/80040234.pdf
PB Springer-Verlag
DT Journal; (online computer file)
LA English
AB DNA translation to the **protein** sequences dets. the common usage of gene name as the enzyme identifier. The previously constructed single-family-member phylogenetic trees are produced by the pair **alignment**. The **alignments** strictly depend upon the user-defined parameters and algorithmic peculiarities, such as but not limited to: homol. **matrix**, initial **gap** penalty value and **gap** elongation function. This rises the necessity to create complete clusterization which reflects the **protein** primary

structure relationships. This **protein**-based clusterization should be made using the objective pair **alignment**. The standard dynamic **alignment** procedure is modified in order to discriminate between the suboptimal resulting **scores**. The special function treats the presence of continuous matching n-tuples as a good property of **alignment**. Pair **alignment** is objectified by finding the optimal **gap** penalty, that allows to get the maximal difference in identity between random and relative sequences. The method is applied to the cytochrome P 450 superfamily. Our sample also contained 15 nitric oxide synthases and 30 random sequences. The similarity **matrix**, obtained by objective pair **alignment**, is worked up by standard UPGMA method.

L1 ANSWER 13 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1998:2346 CAPLUS
DN 128:85617
TI Multiple **alignment** of amino acid sequences using a genetic algorithm
AU Isokawa, Masamichi; Wayama, Masato; Shimizu, Toshio
CS Dep. Information Science, Fac. Sci., Hirosaki Univ., Hirosaki, Japan
SO Science Reports of the Hirosaki University (1997), 44(1), 125-140
CODEN: HUSRAK; ISSN: 0367-6439
PB Hirosaki University, Faculty of Science
DT Journal
LA Japanese
AB We applied a genetic algorithm to the problem of multiple **alignment** of amino acid sequences based on Goldbergs simple genetic algorithm. A sequence including a **gaps** in an **alignment** is represented as a bit string which consists of '0' and '1'. In this bit string, '1' corresponds to a **gap**, with the total number of '0' being exactly the same as the sequence length. The **alignment** is expressed with a **matrix**, which is a vertical arrangement of the bit strings. Bit **matrixes** are prepared as a starting population in a random way: an element in each bit **matrix** is randomly determined to be '0' or '1'. The next population is generated by applying three kinds of genetic operations: crossover, mutation and reproduction The reproduction operation creates the next population from the **matrixes** of the starting population with the use of the ranking selection and the similarity **scores** between amino acids. Next, a window-frame crossover operation exchanges the information partly between two parent **matrixes** selected randomly to produce two child **matrixes**: the amino acid residue correspondes are conserved strictly in this operation. Then, 4 mutation operations ("continuous-**gap**-shift mutation", "continuous-**gap** -extension mutation", "**gap**-block-extension mutation" and " **gap**-block-shift mutation") are applied to bit **matrixes** of the second population. These procedures described above are carried out repeatedly to obtain a nearly optimal **alignment**. We prepared two test data sets of 4 and 5 amino acid sequences from the data base of SWISS-PROT release 30. The amino acid sequences of each data set were aligned with the procedure described above. Nearly optimal **alignments** are obtained by our method. The **alignment** results are comparable to those by ClUSTAL W which is the typical software package for multiple sequence **alignment** base on the tree-based algorithm.

L1 ANSWER 14 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1997:698578 CAPLUS
DN 128:31574
TI Do aligned sequences share the same fold?
AU Abagyan, Ruben A.; Batalov, Serge
CS Biochemistry Department, NYU Medical Center, The Skirball Institute of Biomolecular Medicine, New York, NY, 10016, USA

SO Journal of Molecular Biology (1997), 273(1), 355-368
CODEN: JMOBAK; ISSN: 0022-2836
PB Academic
DT Journal
LA English
AB Sequence comparison remains a powerful tool to assess the structural relatedness of two **proteins**. To develop a sensitive sequence-based procedure for fold recognition, we performed an exhaustive global **alignment** (with zero end **gap** penalties) between sequences of **protein** domains with known three-dimensional folds. The subset of 1.3 million **alignments** between sequences of structurally unrelated domains was used to derive a set of anal. functions that represent the probability of structural significance for any sequence **alignment** at a given sequence identity, sequence similarity and **alignment score**. Anal. of overlap between structurally significant and insignificant **alignments** shows that sequence identity and sequence similarity measures are poor indicators of structural relatedness in the "twilight zone", while the **alignment score** allows much better discrimination between **alignments** of structurally related and unrelated sequences for a wide variety of **alignment** settings. A fold recognition benchmark was used to compare eight different substitution **matrixes** with eight sets of **gap** penalties. The best performing **matrixes** were Gonnet and Blosum50 with normalized **gap** penalties of 2.4/0.15 and 2.0/0.15, resp., while the pos. **matrixes** were the worst performers. The derived functions and parameters can be used for fold recognition via a multilink chain of probability weighted pairwise sequence **alignments**.
RE.CNT 36 THERE ARE 36 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 15 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1997:1458 CAPLUS
DN 126:128946
TI Significant improvement in accuracy of multiple **protein** sequence **alignments** by iterative refinement as assessed by reference to structural **alignments**
AU Gotoh, Osamu
CS Dep. of Biochemistry, Saitama Cancer Center Res. Inst., Saitama, 362, Japan
SO Journal of Molecular Biology (1996), 264(4), 823-838
CODEN: JMOBAK; ISSN: 0022-2836
PB Academic
DT Journal
LA English
AB The relative performances of four strategies for aligning a large number of **protein** sequences were assessed by referring to corresponding structural **alignments** of 54 independent families. Multiple sequence **alignment** of a family was constructed by a given method from the sequences of known structures and their homologues, and the subset consisting of the sequences of known structures was extracted from the whole **alignment** and compared with the structural counterpart in a residue-to-residue fashion. **Gap**-opening and -extension penalties were optimized for each family and method. Each of the four multiple **alignment** methods gave significantly more accurate **alignments** than the conventional pairwise method. In addition, a clear difference in performance was detected among three of the four multiple **alignment** methods examined The currently most popular progressive method ranked worst among the four, and the randomized iterative strategy that optimizes the sum-of-pairs **score** ranked next worst. The two best-performing strategies, one of which was newly developed, both pursue an optimal weighted sum-of-pairs **score**, where the pair wts. were introduced to correct for uneven representations of subgroups in a family. The new method uses doubly nested iterations to

make **alignment**, phylogenetic tree and pair wts. mutually consistent. Most importantly, the improvement in accuracy of **alignments** obtained by these iterative methods over pairwise or progressive method tends to increase with decreasing average sequence identity, implying that iterative refinement is more effective for the generally difficult **alignment** of remotely related sequences. Four well-known amino acid substitution **matrixes** were also tested in combination with the various methods. However, the effects of substitution **matrixes** were found to be minor in the framework of multiple **alignment**, and the same order of relative performance of the **alignment** methods was observed with any of the **matrixes**.

L1 ANSWER 16 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1995:878359 CAPLUS
DN 124:48556
TI **Alignment** of 700 globin sequences: extent of amino acid substitution and its correlation with variation in volume
AU Kapp, Oscar H.; Moens, Luc; Vanfleteren, Jaak; Trotman, Clive N. A.; Suzuki, Tomohiko; Vinogradov, Serge N.
CS Dep. of Radiology and Enrico Fermi Inst., Univ. of Chicago, Chicago, IL, 60637, USA
SO Protein Science (1995), 4(10), 2179-90
CODEN: PRCIEI; ISSN: 0961-8368
PB Cambridge University Press
DT Journal
LA English
AB Seven-hundred globin sequences, including 146 nonvertebrate sequences, were aligned on the basis of conservation of secondary structure and the avoidance of **gap** penalties. Of the 182 positions needed to accommodate all the globin sequences, only 84 are common to all, including the absolutely conserved PheCD1 and HisF8. The mean number of amino acid substitutions per position ranges from 8 to 13 for all globins and 5 to 9 for internal positions. Although the total sequence vols. have a variation .apprx. 2-3%, the variation in volume per position ranges from .apprx.13% for the internal to .apprx.21% for the surface positions. Plausible correlations exist between amino acid substitution and the variation in volume per position for the 84 common and the internal but not the surface positions. The amino acid substitution **matrix** derived from the 84 common positions was used to evaluate sequence similarity within the globins and phycocyanins C and colicins A, via calcn. of pairwise similarity **scores**. The **scores** for globin-globin comparisons over the 84 common positions overlap the globin-phycocyanin and globin-colicin **scores**, with the former being intermediate. For the subset of internal positions, overlap is minimal between the three groups of **scores**. These results imply a continuum of amino acid sequences able to assume the common three-on-three α-helical structure and suggest that the determinants of the latter include sites other than those inaccessible to solvent.

L1 ANSWER 17 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1995:656872 CAPLUS
DN 123:308878
TI An assessment of amino acid exchange **matrixes** in aligning **protein** sequences: the twilight zone revisited
AU Vogt, Gerhard; Etzold, Thure; Argos, Patrick
CS European Mol. Biol. Lab., Heidelberg, D-69126, Germany
SO Journal of Molecular Biology (1995), 249(4), 816-31
CODEN: JMOBAK; ISSN: 0022-2836
PB Academic
DT Journal
LA English
AB The sensitivity of most **protein** sequence **alignment** methods depends strongly on the quality of the comparison **matrixes**

used. These **matrixes**, which assign wts. or similarity **scores** to every possible amino acid substitution pair, are utilized to differentiate amongst the various possible **alignments** of two or more sequences. There are many ways to generate these exchange wts. and new **matrixes** are constantly published. There has been no overall assessment of these various **matrixes** when applied in different **alignment** techniques and over many **protein** folds and families, both close and distant and with the use of several **gap** penalty values. In this work, a set of amino acid sequences matched by superposition of known **protein** tertiary topologies is used to test the **alignment** accuracy of the different method/ **matrix**/penalty combinations. The comparisons show relatively similar results for the top scoring **matrixes**, a preference for the global **alignment** method of Needleman and Wunsch, and the importance of **matrix** modification and optimized **gap** penalties. The relationship between the percentage identity in a resulting **alignment** and the level of correctness to be expected are given for the top-performing **matrix**, resulting in a better definition of the so-called "twilight zone". Ests. are made for the probability that two sequences, aligned at a certain level of residue percentage identity, are in fact unrelated.

L1 ANSWER 18 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1994:128703 CAPLUS
DN 120:128703
TI Sequence **alignment** and penalty choice. Review of concepts, case studies and implications
AU Vingron, Martin; Waterman, Michael S.
CS Dep. Math., Univ. South. California, Los Angeles, CA, 90089-1113, USA
SO Journal of Molecular Biology (1994), 235(1), 1-12
CODEN: JMOBAK; ISSN: 0022-2836
DT Journal; General Review
LA English
AB A review with 24 refs. **Alignment** algorithms to compare DNA or amino acid sequences are widely used tools in mol. biol. The algorithms depend on the setting of various parameters, most notably **gap** penalties. The effect that such parameters have on the resulting **alignments** is still poorly understood. This paper begins by reviewing two recent advances in algorithms and probability that enable the authors to take a new approach to this question. The first tool the authors introduce is a newly developed method to delineate efficiently all optimal **alignments** arising under all choices of parameters. The second tool comprises insights into the statistical behavior of optimal **alignment scores**. From this the authors gain a better understanding of the dependence of **alignments** on parameters in general. The authors propose novel criteria to detect biol. good **alignments** and highlight some specific features about the interaction between similarity **matrixes** and **gap** penalties. to illustrate the authors' anal. the authors present a detailed study of the comparison of two Ig sequences.

L1 ANSWER 19 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1993:665894 CAPLUS
DN 119:265894
TI **Protein** structure comparison by **alignment** of distance **matrixes**
AU Holm, Liisa; Sander, Chris
CS Protein Des. Group, Eur. Mol. Biol. Lab., Heidelberg, D-69012, Germany
SO Journal of Molecular Biology (1993), 233(1), 123-38
CODEN: JMOBAK; ISSN: 0022-2836
DT Journal
LA English
AB With a rapidly growing pool of known tertiary structures, the importance of **protein** structure comparison parallels that of sequence

alignment. The authors have developed a novel algorithm (Dali) for optimal pairwise **alignment** of **protein** structures. The three-dimensional coordinates of each **protein** are used to calculate residue-residue (Cα-Cα) distance **matrixes**. The distance **matrixes** are first decomposed into elementary contact patterns, e.g., hexapeptide-hexapeptide submatrixes. Then, similar contact patterns in the two **matrixes** are paired and combined into larger consistent sets of pairs. A Monte Carlo procedure is used to optimize a similarity **score** defined in terms of equivalent intramol. distances. Several **alignments** are optimized in parallel, leading to simultaneous detection of the best, second-best and so on solns. The method allows sequence **gaps** of any length, reversal of chain direction and free topol. connectivity of aligned segments. Sequential connectivity can be imposed as an option. The method is fully automatic and identifies structural resemblances and common structural cores accurately and sensitively, even in the presence of geometrical distortions. An all-against-all **alignment** of over 200 representative **protein** structures results in an objective classification of known three-dimensional folds in agreement with visual classifications. Unexpected topol. similarities of biol. interest have been detected, e.g., between the bacterial toxin colicin A and globins and between the eukaryotic POU-specific DNA-binding domain and the bacterial λ repressor.

L1 ANSWER 20 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1993:442704 CAPLUS
DN 119:42704
TI PROFALIGN: a computer program to graphically align biological sequences
AU Ochagavia, M.; Ricardo, R.; Fernandez de Cossio, J.; Bringas, R.
CS Cent. Ing. Genet. Biotecnol., Havana, Cuba
SO Biotecnologia Aplicada (1992), 9(2), 174-9
CODEN: BTAPEP; ISSN: 0864-4551
DT Journal
LA Spanish
AB A computer program to calculate and graphically show the **alignment** profile of two biol. sequences is described. The program produces an **alignment** profile which is calculated using a previous two sequences **alignment**, a weight **matrix** and a **gap** penalty. The function which describes the profile is evaluated for each position taking into account the similarity **score** of its neighbor positions. This program is useful to find conserved regions and to evaluate the similarity level of two sequences in every region.

L1 ANSWER 21 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1993:76518 CAPLUS
DN 118:76518
TI MATCH-BOX: a fundamentally new algorithm for the simultaneous **alignment** of several **protein** sequences
AU Depiereux, Eric; Feytmans, Ernest
CS Dep. Biol., Fac. Univ. Notre-Dame Paix, Namur, B-5000, Belg.
SO CABIOS, Computer Applications in the Biosciences (1992), 8(5), 501-9
CODEN: COABER; ISSN: 0266-7061
DT Journal
LA English
AB Original algorithms for simultaneous **alignment** of **protein** sequences are presented, including sequence clustering and within- or between-groups multiple **alignment**. The way of matching similar regions is fundamentally new. Complete matches are formed by segments more similar than expected by random, according to a given probability limit. Any classic or user-defined **score** **matrix** can be used to express the similarity between the residues. The algorithm seeks for complete matches common to all the sequences without performing pairwise **alignment** and regardless of **gap** weighting. An automatic screening delineates all the similar

regions (boxes) that may be defined for a given maximal shift between the sequences. The shift can be large enough to allow the matching of any region of a sequence with any region of another one. It can also be short and used to refine the **alignment** around anchor points. The algorithm provides the most likely optimal **alignment** and a comprehensive list of the **alignment** dilemma. Duality between automatism and interactivity is provided. Depending on the problem complexity, a final **alignment** is obtained fully automatically or requires some interactive handling to discriminate alternative pathways.

L1 ANSWER 22 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1993:2680 CAPLUS
DN 118:2680
TI Local multiple **alignment** by consensus **matrix**
AU Alexandrov, Nickolai N.
CS Fac. Sci., Kyoto Univ., Kyoto, 606, Japan
SO CABIOS, Computer Applications in the Biosciences (1992), 8(4), 339-45
CODEN: COABER; ISSN: 0266-7061
DT Journal
LA English
AB A new algorithm for aligning several sequences based on the calcn. of a consensus **matrix** and the comparison of all the sequences using this consensus **matrix** is described. This consensus **matrix** contains the preference **scores** of each nucleotide/amino acid and **gaps** in every position of the **alignment**. Two modifications of the algorithm corresponding to the evolutionary and functional meanings of the **alignment** were developed. The first one solves the best-fitting problem without any penalty for end **gaps** and with an internal **gap** penalty function independent of the **gap** length. This algorithm should be used when comparing evolutionary-related **proteins** for identifying the most conservative residues. The other modification of the algorithm finds the most similar segments in the given sequences. It can be used for finding those parts of the sequences that are responsible for the same biol. function. In this case, the **gap** penalty function was chosen to be proportional to the **gap** length. The result of aligning amino acid sequences of neutral proteases and a compilation of 65 allosteric effectors and substrates of PEP carboxylase are presented.

L1 ANSWER 23 OF 23 CAPLUS COPYRIGHT 2004 ACS on STN
AN 1989:402931 CAPLUS
DN 111:2931
TI A multiple sequence **alignment** algorithm for homologous **proteins** using secondary structure information and optionally keying **alignments** to functionally important sites
AU Henneke, Christina M.
CS Sch. Chem., Univ. Bath, Bath, BA2 7AY, UK
SO CABIOS, Computer Applications in the Biosciences (1989), 5(2), 141-50
CODEN: COABER; ISSN: 0266-7061
DT Journal
LA English
AB The programs described herein function as part of a suite of programs designed for pairwise **alignment**, multiple **alignment**, generation of randomized sequences, production of **alignment** **scores**, and a sorting routine for anal. of the **alignments** produced. The sequence **alignment** programs penalize **gaps** (absences of residues) within regions of **protein** secondary structure and have the added option of fingerprinting structurally or functionally important **protein** residues. The multiple **alignment** program is based upon the sequence **alignment** method of Needleman and Wunsch and the multiple **alignment** extension of Barton and Sternberg. Application includes the feature of optionally weighting active site, monomer-monomer, ligand contact, or other important template residues to bias the **alignment** toward

matching these residues. A sum-**score** for the **alignments** is introduced, which is independent of **gap** penalties. This **score** more adequately reflects the character of the **alignments** for a given scoring **matrix** than the **gap**-penalty-dependent total **score** described previously in the literature. In addition, individual amino acid similarity **scores** at each residue position in the **alignments** are printed with the **alignment** output to enable immediate quant. assessment of homol. at key sections of the aligned chains.

FILE 'USPATFULL' ENTERED AT 14:52:42 ON 04 MAY 2004
L3 4 S L1/CLM

=> d bib,abs,kwic 1-4

L3 ANSWER 1 OF 4 USPATFULL on STN
AN 2004:19991 USPATFULL
TI Multiple sequence alignment
IN Swindells, Mark, Lincs, UNITED KINGDOM
Rae, Mark, London, UNITED KINGDOM
PI US 2004015298 A1 20040122
AI US 2002-221833 A1 20021219 (10)
WO 2001-GB1110 20010314
DT Utility
FS APPLICATION
LREP DARBY & DARBY P.C., P. O. BOX 5257, NEW YORK, NY, 10150-5257
CLMN Number of Claims: 16
ECL Exemplary Claim: 1
DRWN 3 Drawing Page(s)
LN.CNT 627
CAS INDEXING IS AVAILABLE FOR THIS PATENT.
AB The invention relates to a method of aligning a plurality of sequences. In a similar way to known multiple alignment methods, the method of the invention uses a profile for the nominated sequence in an alignment strategy. The key novel concept behind the method of the invention is to allow the profile to be extended in regions where gaps are desired. This alternative strategy is implemented using pre-generated profiles as a basis for the multiple alignment.

CAS INDEXING IS AVAILABLE FOR THIS PATENT.
CLM What is claimed is:
1. A computer-implemented method of aligning a plurality of **protein** or nucleic acid sequences comprising the steps of: a) performing an **alignment** of a query sequence to a target sequence using a dynamic programming algorithm that constructs the **alignment** using a scoring **matrix** profile to provide an **alignment score** for aligning amino acid residues together, wherein suitable candidate residues for **alignment** are given a positive **score** and unsuitable candidate residues are given a negative **score**, and negative **score** penalties are generated both for opening and for extending a **gap** in one of the sequences in the **alignment**; and b) repeating step a) for each sequence to be aligned; wherein the scoring **matrix** profile is modified after each **alignment** step a) and before being used to generate the **alignment** of the next sequence, and wherein if the best scoring **alignment** requires that a **gap** be introduced into the profile, the profile is modified by inserting the residues from the query sequence that match up with the **gap** region.

. . . the profile where residues or nucleotides have been inserted and said amino acid residues or nucleotides are assigned a negative **score**, their **score** is reset to zero, such that multiple sequences that have similar regions that were not present in the original profile may be aligned together without penalty while at the same time allowing the **alignment score** to be increased for correctly aligned regions that have a positive **score**.

3. A method according to either claim 1 or claim 2, wherein if the **alignment** of a second or subsequent query sequence requires that a **gap** be inserted or extended into the sequence that is being aligned against the profile and this **gap** falls within a modified region of the profile where residues or nucleotides have been inserted, no negative **score** penalty is generated, such that

sequence that would normally align against the profile without the need for a **gap** can be aligned without an inserted region interfering with the **alignment**.

. . . claims, wherein if a query sequence is known to align against a target sequence in multiple locations such that multiple **alignment** hits are generated by the **alignment** of these sequences, then step a) is repeated for each location at which the sequences align, and for each separate iteration, the **alignment** of the sequences is constrained to one particular **alignment** location.

5. A method according to claim 4, wherein the **alignment** is constrained by excluding regions from consideration by the dynamic programming algorithm by setting the **matrix** profile **scores** in the excluded region to a large negative value beyond a value that would occur naturally during the execution of. . .

. . . the large negative value assigned is the largest negative value that can be stored by the computer on which the **alignment** method is being performed.

7. A method according to any one of the preceding claims, wherein the scoring **matrix** profile that is used in the **alignment** method is a profile generated by running a profile-based **alignment** algorithm on the target sequence.

8. A method according to claim 7, wherein the profile-based **alignment** algorithm is the position specific iterated basic local **alignment** search tool (PSI-BLAST).

9. A method according to any one of claims 1-7, wherein the scoring **matrix** profile that is used in the **alignment** method is a default scoring **matrix**.

10. A method according to claim 9, wherein said default **matrix** is a BLOSUM or PAM **matrix**.

12. A computer apparatus according to claim 11 comprising: a processor means comprising: a memory means adapted for storing data relating to amino acid or nucleotide sequences; means for inputting data relating to a plurality of **protein** or nucleic acid sequences; computer software means stored in said computer memory adapted to align said plurality of **protein** or nucleic acid sequences and output a multiple **alignment** of said sequences.

13. A computer-based system for aligning a plurality of **protein** or nucleic acid sequences comprising: means for inputting data relating to a plurality of **protein** or nucleic acid sequences; means adapted to align said plurality of **protein** or nucleic acid sequences; and means for outputting a multiple **alignment** of said sequences.

14. A system according to claim 13, wherein said means adapted to align said plurality of **protein** or nucleic acid sequences is a computer software means.

. . . device; the memory storing a module that is configured so that upon receiving a request to align a plurality of **protein** or nucleic acid sequences, it performs the steps listed in any one of claims 1-10.

. . . computer program mechanism comprising a module that is configured so that upon receiving a request to align a plurality of **protein** or nucleic acid sequences, it performs the steps listed in any one of claims 1-10.

L3 ANSWER 2 OF 4 USPATFULL on STN
AN 2003:266568 USPATFULL
TI Database
IN Swindells, Mark, Easton-on- the- Hill, UNITED KINGDOM
Thornton, Janet, Herts, UNITED KINGDOM
Jones, David, London, UNITED KINGDOM
PI US 2003187587 A1 20031002
AI US 2003-221831 A1 20030204 (10)
WO 2001-GB1105 20010314
PRAI GB 2000-6153 20000314
DT Utility
FS APPLICATION
LREP DARBY & DARBY P.C., P. O. BOX 5257, NEW YORK, NY, 10150-5257
CLMN Number of Claims: 57
ECL Exemplary Claim: 1
DRWN 16 Drawing Page(s)
LN.CNT 3748
AB The invention concerns methods and systems for predicting the function of proteins. In particular, the invention relates to databases in which details of sequence homologies, biological functions and structures that are shared between proteins of differing sequence have been compiled. The invention also relates to methods, systems and computer software that allows the prediction of protein function and structure and, optionally, the ligand binding properties of the proteins within such a database.

CLM What is claimed is:
1) A method of compiling a database containing information relating to the interrelationships between different **protein** and/or nucleic acid sequences, said method comprising the steps of: a) integrating data from one or more separate sequence data. . . comparing each query sequence in the combined database with the other sequences represented in the combined database to identify homologous **proteins** or nucleic acid sequences; c) compiling the results of the comparisons generated in step b) into a database; and d). . .
2) A method of compiling a database containing information relating to the interrelationships between different **protein** sequences, said method comprising the steps of: a) integrating **protein** data from one or more separate sequence data resources and one or more structural data resources into a combined database; b) comparing each query **protein** sequence in the combined database with the other **protein** sequences represented in the combined database to identify homologous **proteins** using, for each query sequence: i) one or more pairwise sequence **alignment** searches, ii) one or more profile-based sequence **alignment** searches; iii) one or more threading-based approaches; c) compiling the results of the comparisons generated in step b) into a. . .
5) A method according to either claim 2 or claim 3, wherein said structural data resource is the **Protein** Data Base (PDB).

. . . A method according to any one of the preceding claims, wherein said integrating step (a) includes the step of scanning **protein** sequences against regular expressions and profiles recorded in a database that contains information relating to annotations of sequence families and. . .
10) A method according to claim 9, wherein **protein** sequences are scanned against regular expressions and profiles in the PROSITE database.

16) A method according to claim 15, wherein the **alignment** of each sequence with the longest sequence in its group is specified by indexing the start and the end points of the sequence **alignment**

19) A method according to claim 18, wherein said compositionally-biased regions are selected from one or more of signal **peptides**, coiled-coil regions, membrane regions, and other regions of low complexity.

20) A method according to claim 19, wherein signal **peptides**, coiled-coil regions, membrane regions, and regions of low complexity are masked for exclusion in comparison step (b).

21) A method according to any one of the preceding claims, wherein said comparison step (b)(i) comprises a pairwise **alignment** search in which each selected sequence in the database generated in step (a) is compared against each other selected sequence.

. . .

22) A method according to claim 21, wherein said comparison step (b)(i) is performed using a gapped BLAST sequence **alignment** algorithm.

. . . to any one of claims 20-22, wherein a sequence profile relating to position-specific substitution probabilities is generated from the pairwise **alignment** search if a significant number of hits are found between sequences in the database and the query sequence to allow.

. .

. . . claim 23, wherein for each sequence in the composite database, the profile generated by the final iteration of the pairwise **alignment** search is selected as the profile for use in the profile-based **alignment** search, and wherein for sequences in the collated database against which too few sequences aligned to allow the generation of a meaningful profile, a substitution **matrix** is used as a default profile.

25) A method according to claim 24, wherein said substitution **matrix** is the BLOSUM62 **matrix** or PAM 250 **matrix**.

26) A method according to any one of the preceding claims, wherein a PSI-BLAST-based search is used for the profile-based **alignment** search of step (bii).

27) A method according to claim 24-26, wherein in the profile-based **alignment** search, for each target sequence, identified hits are clustered according to sequence hit, and the clustered sequences are checked for. . . wherein significant overlap is assessed using a graph subset construction algorithm, such that duplicated or redundant information generated in the **alignment** is reduced.

30) A method according to any one of the preceding claims, wherein multiple **alignments** are generated of sequences in the database.

31) A method according to claim 32, wherein each multiple **alignment** comprises the steps of: a) performing a pairwise **alignment** of a query sequence to a target sequence using a dynamic programming algorithm that constructs the **alignment** using a scoring **matrix** profile to provide an **alignment** **score** for aligning amino acid residues together, wherein suitable candidate residues for **alignment** are given a positive **score** and unsuitable candidate residues are given a negative **score**, and negative **score** penalties are generated for both opening and extending a **gap** in one of the sequences in the **alignment**; and b) repeating step a) for each sequence to be aligned; wherein the scoring **matrix** profile is modified

after each **alignment** step and before being used to generate the **alignment** of the next sequence to be aligned.

32) A method according to claim 31, wherein if the best scoring **alignment** requires that a **gap** be introduced into the profile, the profile is modified by inserting the residues from the query sequence that match up with the **gap** region.

. . . the profile where residues or nucleotides have been inserted and said amino acid residues or nucleotides are assigned a negative **score** , their **score** is reset to zero, such that multiple sequences that have similar regions that were not present in the original profile may be aligned together without penalty while at the same time allowing the **alignment score** to be increased for correctly aligned regions that have a positive **score**.

34) A method according to any one of claims 31-33, wherein if the **alignment** of a second or subsequent query sequence requires that a **gap** be inserted or extended into the sequence that is being aligned against the profile and this **gap** falls within a modified region of the profile where residues or nucleotides have been inserted, no negative **score** penalty is generated, such that sequence that would normally align against the profile without the need for a **gap** can be aligned without an inserted region interfering with the **alignment**.

. . . 31-34, wherein if a query sequence is known to align against a target sequence in multiple locations such that multiple **alignment** hits are generated by the **alignment** of these sequences, then step a) is repeated for each location at which the sequences align, and for each separate iteration, the **alignment** of the sequences is constrained to one particular **alignment** location.

36) A method according to claim any one of claims 31-35, wherein the **alignment** is constrained by excluding regions from consideration by the dynamic programming algorithm by setting the **matrix** profile **scores** in the excluded region to a large negative value beyond a value that would occur naturally during the execution of.
. .
. . . the large negative value assigned is the largest negative value that can be stored by the computer on which the **alignment** method is being performed.

38) A method according to any one of claims 31-37, wherein the results of the **alignment** are loaded into the database.

39) A method according to any one of the preceding claims, wherein in said comparison step (biii), a pairwise **alignment** is performed between a query sequence of unknown structure and a sequence of known structure, followed by a structure overlay step in which the generated **alignment** is used to match a structure to the query sequence of unknown structure.

40) A method according to claim 39, wherein the pairwise **alignment** has two modes: a forward mode, in which the profile for the sequence of known structure is used to identify areas of **alignment** with the query sequence; and a reverse mode, in which the profile for the query sequence of unknown structure is used to identify areas of **alignment** with the sequence of known structure, such that a proposed **alignment** and confidence value are output for each pairwise **alignment**.
41) A method according to either claim 39 or claim 40, wherein both a local and global pairwise **alignment** is performed.

42) A method according to claim 41, wherein said local **alignment** utilises the Smith-Waterman algorithm and said global **alignment** utilises a Myers-Miller-based algorithm.

. . . step comprises the steps of: a) overlaying the residues of the known structure with the corresponding residues from the pairwise **alignment** in the sequence of unknown structure; b) summing the accessibility potential for each residue to give a total accessibility **score**; c) summing the pairwise contributions from each residue-residue interaction for each of the atom pairs to give a total pairwise energy value; d) inserting the total accessibility **score**, total pairwise energy value and **alignment** **score** into a neural network that combines these three values into a single **score**; and e) comparing this single **score** to a value calculated for a training set based on a selection of relationships from all of the possible combinations. . . structures to give a confidence value that reflects the percentage probability of a relationship being correct for a given network **score**.

46) A database containing information relating to the degree of similarity/interrelationships between different **protein** sequences generated by a method, system or apparatus according to any one of the preceding claims.

47) A database system comprising: a database of **protein** or nucleic acid sequence entries containing sequence information, optionally structure information, functional annotation, and information relating to the **alignment** of each sequence in the database with every other sequence in the database; a plurality of computer programs for processing. . .

49) A computer apparatus for compiling a database containing information relating to the similarity between different **proteins**, said apparatus comprising: a processor means comprising: a memory means adapted for storing data relating to amino acid sequences and the relationships shared between different **protein** sequences; first computer software stored in said computer memory adapted to align said **protein** sequences using one or more pairwise **alignment** approaches; second computer software stored in said computer memory adapted to align said **protein** sequences using one or more profile-based approaches; third computer software stored in said computer memory adapted to align said **protein** sequences using one or more threading-based approaches.

50) A computer apparatus according to claim 49, wherein said memory means is adapted for storing data relating to: (a) the sequences of a plurality of **proteins** or nucleic acids; (b) the structures of a plurality of **proteins**; (c) the predicted **alignments** of each of said sequences with every other one of said sequences; (d) the predicted **alignments** of sequences of known structure with those of unknown structure; (e) annotation of the sequences.

51) A computer apparatus for predicting the biological function of a **protein** comprising: a processor means comprising: a computer memory for storing a specific sequence of amino acid residues; first computer software. . . application programming interface; display means, connected to said processor for visually displaying to a user on command a list of **proteins** with which said specific sequence of amino acid residues is predicted to share a biological function.

52) A computer system for compiling a database containing information relating to the similarity between different **protein** or nucleic acid sequences, said system performing the steps of: a) combining sequence data from separate sequence data resources into. .

. comparing each query sequence in the composite database with the other sequences represented in the composite database to identify homologous **proteins** or nucleic acids using, for each query sequence: i. one or more pairwise sequence **alignment** searches, ii. one or more profile-based sequence **alignment** searches; iii. optionally, one or more threading-based approaches; c) outputting the results of the comparisons generated in step b) into. . .

53) A computer-based system for predicting the biological function of a **protein** comprising the steps of: a) inputting a query sequence of amino acids whose function is to be predicted into a. . .

54) A computer-based system for predicting the biological function of a **protein** comprising the steps of: a) accessing a database according to claim 46 or claim 47, b) inputting a query sequence. . .

55) A computer system for predicting the biological function of a **protein**, comprising: a central processing unit; an input device for inputting requests; an output device; a memory; at least one bus. . . memory storing a module that is configured so that upon receiving a request to predict the biological function of a **protein**, it performs the steps listed in any one of claims 1-45.

56) A computer-based method for predicting the biological function of a **protein**, comprising the steps of: a) accessing the database of claim 46 or 47, at a remote site, b) inputting into. . .

. . . mechanism comprising a module that is configured so that upon receiving a request to predict the biological function of a **protein**, it performs a method as recited in any one of claims 1-45.

L3 ANSWER 3 OF 4 USPATFULL on STN
AN 2003:147249 USPATFULL
TI Methods for establishing a pathways database and performing pathway searches
IN Yang, Yonghong, San Jose, CA, UNITED STATES
Tillinghast, John, Cupertino, CA, UNITED STATES
Piercy, Christopher, Cupertino, CA, UNITED STATES
PA Genmetrics (U.S. corporation)
PI US 2003100996 A1 20030529
AI US 2002-81904 A1 20020220 (10)
PRAI US 2002-347019P 20020107 (60)
US 2001-269711P 20010220 (60)
DT Utility
FS APPLICATION
LREP Genmetrics, Inc., 4230 Ranwick Ct., San Jose, CA, 95118
CLMN Number of Claims: 22
ECL Exemplary Claim: 1
DRWN 11 Drawing Page(s)
LN.CNT 1868
CAS INDEXING IS AVAILABLE FOR THIS PATENT.
AB The invention provides a computerized storage and retrieval system for storing biological information organized as a protein pathways database and methods for performing pathway searches on nodes (proteins or other molecules), modes (interactions), and nodes-and-modes. The protein pathways database is a relational database that integrates protein sequence, genomic sequence, gene-expression, protein interactions, protein-protein association and pathway data and can be searched using a query pathway to predict homologous or orthologous nodes, modes, and pathways.

CAS INDEXING IS AVAILABLE FOR THIS PATENT.
CLM What is claimed is:
. . . means for displaying the data; a programmable central processing unit for performing automated analysis; and a data storage means

containing **protein** pathways and annotated information on the pathways stored in a relational database, wherein the pathways annotated and organized in a. . .

. . . system of claim 1, wherein the information pertaining to the pathways is stored in a plurality of tables further comprising **proteins**, their sequences and attributes; **protein** interactions; **protein-protein** associations; **protein** pathways; mRNA, microarray, and **protein** expression data; genes, their sequences and attributes; and descriptions of cells, tissues, organs, pathology reports, patient histories, and treatments.

. . . the central processing unit is programmed to retrieve, input, edit, annotate, search, calculate similarities, align, and predict homologous or orthologous **protein** pathways.

4. The computer system of claim 1, wherein the central processing unit is programmed to perform **protein** sequence analysis, **protein** interactions analysis, **protein-protein** association analysis, **protein** pathway analysis, gene expression analysis, pathway annotation analysis, pathway edit analysis, pathway expression analysis, tissue expression analysis, subtractive hybridization analysis,. . .

. . . a means for displaying the data is used to show two related pathways as a diagram containing nodes which represent **proteins** or non-**protein** molecules; modes that represent **protein** interactions or **protein-protein** associations; **scores** calculated from sequence, motif or structural homologies that interrelate nodes; and coefficients of similarity that interrelate modes of the pathway.

. . .

7. The computer system of claim 1, wherein the central processing unit is programmed to compare two **protein** pathways by a node-only, a mode-only, or a node-and-mode comparison and wherein the node-only comparison is selected from **protein** only, non-**protein** only, and **protein** and non-**protein** nodes.

. . . two nodes, defined as ##EQU9## c) using traceback to identify putative pathways PPW.sub.j, 1<=j<=max n.sub.i with the top n best **scores**.

. . . 9. A method for performing pathway editing comprising programming the central processing unit of claim 1 to identify interactions among **proteins**; weigh the interactions; and calculate coefficients of similarity for the interactions, thereby producing an OS **score** and editing the **protein** pathway.

10. A method of using genes which encode known **proteins** to annotate modes of a **protein** pathway comprising: a) using the computer system of claim 1 to select genes which encode known **proteins**, b) employing the genes to produce a **protein**-**protein** association **matrix** containing coefficients of similarity, and c) annotating the modes of the pathway using the coefficients of similarity from the **matrix**.

11. A method for **protein** pathways analysis using a node-and-mode comparison comprising: a) submitting a query pathway and **protein** sequences; and b) allowing the computer system of claim 1 to i) compare nodes using the dynamic programming algorithm wherein a sequence identity **score** or p-value summarizes similarity and wherein a weighting factor between 0 and 1 is assigned to corresponding nodes, ii) compare modes by generating a SCIM **matrix**, thereby assigning a coefficient of similarity to corresponding modes, iii) align pathways globally or locally, wherein insertion or deletion of nodes or modes incurs a penalty, iv) sum all similarity **scores**, and v) display at least one high-scoring segment of the aligned

pathways.

12. A method for performing **protein** pathways analysis comprising: a) submitting a query pathway and **protein** sequences; and b) allowing the computer system of claim 1 to i) organize and analyze the query pathway and **protein** sequences, ii) compare **protein** sequence identity of the query with all **protein** sequences in the **protein** pathways database using standard methods of **protein** comparison, iii) use a SCIM **matrix** to derive and compare coefficients of similarity for each interaction of the query and all interactions for **proteins** in the **protein** pathways database, iv) calculate an OS-**score** based on sequence identity and coefficients of similarity, remove all pathways not meeting user-specified threshold for OS-**score**, and vi) retrieve aligned pathways meeting the threshold.

13. A method for searching a **protein** pathways database for **protein** interactions comprising: a) submitting a query pathway; b) allowing the central processing unit of claim 1 to perform **protein** interactions analysis between the query pathway and all **protein** pathways in the **protein** pathways database wherein coefficient of similarity is produced to interrelate each mode of the query pathway and a mode of the most closely related **protein** pathway; and c) retrieving at least one **protein** pathway **alignment**.

14. A method of using a query pathway to search a **protein** pathways database to predict homologous pathways comprising: a) submitting a query pathway and **protein** sequences; b) allowing the central processing unit of claim 1 to compare the query pathway and **protein** sequences with all **protein** pathways and **proteins** in the **protein** pathways database, and c) retrieving a plurality of pathway **alignments** wherein the homologous pathways are aligned by OS-**score**.

15. A method of using a known **protein** pathway and a **protein** database to predict orthologous pathways comprising: a) submitting a query pathway and known **protein** sequences, b) allowing the central processing unit of claim 1 to compare known sequences to all **protein** sequences stored in the database, c) retrieving orthologous **proteins** with the highest identity to the known **proteins**, d) inheriting **protein** interactions from the query pathway, and e) aligning the query pathway and the orthologous **proteins**, thereby predicting orthologous pathways.

16. A method of using a known **protein** pathway to predict the nodes and modes of a novel pathway comprising: a) submitting a query pathway and known **protein** sequences; b) applying standard methods of comparison to determine similarity between the known **protein** sequences and **protein** sequences in the **protein** databases, thereby predicting candidate nodes; c) utilizing coefficients of similarity from **protein** interactions or **protein-protein** association data, thereby predicting candidate modes; and d) retrieving novel pathways with an OP-**score** obtained using an optimization algorithm.

17. The method of claim 16, wherein coefficients of similarity are based on mRNA/cDNA counting, microarray expression, **protein** expression, known **protein-protein** associations, a promoter similarity **matrix**, or more than one of these methods.

. . . method is average linkage, single linkage, complete linkage, K-means, or self-organizing maps; the constraint is that no more than one

protein in each cluster is derived from a single column of aligned **proteins**; and the accuracy of the prediction is determined by an OP-**score**.

19. A method for predicting novel pathways comprising: a) generating candidate **proteins** from one species for each node based on a **protein** search; b) employing a means for optimization to find likely linear linkages between candidate **proteins** aligned to the query pathway with possible **gaps** in the **alignment** , and c) reporting all pathways with optimal and sub-optimal predictions that satisfy user-specified **alignment** and interaction parameters wherein the accuracy of the prediction is provided by OP-**score**.

21. A method for determining the function of a **protein** or a gene that encodes the **protein** comprising: a) placing the **protein** encoded by the gene in a candidate pathway involving at least two **proteins**, and b) using the data storage means of claim 1 wherein the interactions with **proteins** and non-**protein** molecules, cellular location, and expression are used to determine the function of the **protein** or gene.

22. A method for predicting novel pathways comprising: a) submitting a query pathway and **protein** sequence b) using the computer system of claim 1 to process the query pathway and **protein** sequences using orthologous pathway prediction wherein the data is derived from **protein** similarities and interactions, or homologous pathway prediction wherein the data is derived from **protein** similarities and interactions, from **protein**-**protein** associations, and c) applying a dynamic programming algorithm or a constrained clustering algorithm, thereby predicting the novel pathways.

L3 ANSWER 4 OF 4 USPATFULL on STN
AN 2001:104153 USPATFULL
TI Method of searching database of three-dimensional protein structures
IN Toh, Hiroyuki, Suita, Japan
PA Biomolecular Engineering Research Institute, Suita, Japan (non-U.S. corporation)
PI US 6256647 B1 20010703
AI US 1999-250730 19990216 (9)
PRAI JP 1998-32503 19980216
DT Utility
FS GRANTED
EXNAM Primary Examiner: Choules, Jack; Assistant Examiner: Le, Debbie M.
LREP Oblon, Spivak, McClelland, Maier & Neustadt, P.C.
CLMN Number of Claims: 7
ECL Exemplary Claim: 1
DRWN 35 Drawing Figure(s); 31 Drawing Page(s)
LN.CNT 661
AB A method of searching a database of three-dimensional protein structures. The method comprises the steps of setting a three-dimensional protein structure; forming a two-dimensional binary distance map based on the three-dimensional protein structure; forming a one-dimensional peripheral distribution based on the distance map; and comparing the one-dimensional peripheral distribution of a protein structure with that of another protein structure a dynamic programming algorithm. The method increases detection sensitivity and search speed.

CLM What is claimed is:
1. A method of searching a database of three-dimensional **protein** structures, comprising the steps of: (a) setting a three-dimensional **protein** structure; (b) forming a two-dimensional binary distance

map based on the three-dimensional **protein** structure; (c) forming a one-dimensional peripheral distribution based on the binary distance map; and (d) comparing the one-dimensional peripheral distribution with that for another three-dimensional **protein** structure by a dynamic programming algorithm.

2. A method of searching a database of three-dimensional **protein** structures according to claim 1, wherein said distance map is a two dimensional image and has a structure of a triangular **matrix** in which respective columns or respective rows correspond to respective residues of a **protein**; the i-th row corresponds to the i-th amino acid residue counted from the N terminal end, and the j-th column corresponds to the j-th amino acid residue counted from the N terminal end; each element (i, j) of the **matrix** corresponds to the distance between the a carbon of the i-th residue and the a carbon of the j-th residue;. . .

3. A method of searching a database of three-dimensional **protein** structures according to claim 2, wherein said peripheral distribution is composed of a vertical peripheral distribution obtained as a distribution. . .

4. A method of searching a database of three .dimensional **protein** structures according to claim 3, wherein for comparison between peripheral distributions, an **alignment score** obtained by the dynamic programming algorithm divided by the **alignment** length is used as a similarity between two structures.

5. A method of searching a database of three-dimensional **protein** structures according to claim 3, wherein a two dimensional **matrix**, D, is used for the comparison of peripheral distributions; each element of the **matrix** D is obtained by solving the following recurrence equation; through the solution of the equation, the similarity is accumulated from the upper left corner toward the lower right corner of the **matrix** D, considering insertion and deletion; and then, the similarity between two peripheral distributions is obtained as a value for the element of the lower right of the **matrix** D: D.sub.i,j =max {D.sub.i-1, j-1 +s.sub.i, j, D.sub.i-1, j -g, D.sub.i, j-1 -g} where g=5 : **gap** penalty (however, g=0 at the boundary), and S.sub.i, j is represented by the following equation and indicates the similarity between the i-th element of the peripheral distribution of **protein** A and the j-th element of the peripheral distribution of **protein** B: S.sub.i, j =a/{(N.sup.A.sub.i -N.sup.B.sub.j).sup.2 +b}+a/{(C.sup.A.sub.i -C.sup.B.sub.j).sup.2 +b} where N.sup.A.sub.i indicates the j-th frequency of the vertical peripheral distribution of **protein** A; C.sup.A.sub.i indicates the i-th frequency of the horizontal distribution of **protein** A; N.sup.B.sub.j indicates the j-th frequencies of the vertical peripheral distributions of **protein** B; C.sup.B.sub.j indicates the j-th frequencies of the horizontal peripheral distribution of **protein** B; and a and b are constants.

6. A method of searching a database of three-dimensional **protein** structures according to claim 3, wherein a dot frequency R in the distance map is defined as follows: R=number of. . .

7. A method of searching a database of three dimensional **protein** structures according to claim 3, wherein the threshold is determined such that the dot frequency R falls within the range. . .